CS 441 Distributed Systems in Cloud Computing
Homework 2

Input

The program takes the input of Word2Vec from assignment 1. The Vector embedding is of dimension 100 (E = 100).

Sliding window

The spark cluster is used to parallelize the computation of creating Sliding window from the vector embeddings. We have approximately 30,000 embeddings. We have taken a sequence length of 15 (L = 15). The output of sliding window is a tensor of dimension (2950*15*100).

Neural network

The output of the Sliding window tensor is passed as input to the Neural network for training and the computation is parallelized using spark. The output of training is a trained model (trained_model.zip) and is in the root folder of the project in my case.

Embedding Prediction

In this step we take a series of vector embedding as our input and predict the next word embedding as output of our Large language Model.

Baseline Submission

For the baseline submission we have logged the statistics for Neural network and our Output Embedding vectors we have the following values into our statistics.txt file

1) Training Loss and Accuracy
2) Gradient Norm
3) Learning Rate
4) Memory Usage
5) Time per Epoch/Iteration
6) Data Shuffling and Partitioning Statistics (Attached as part of submission from spark-UI)
7) Spark-Specific Metrics
8) The output predicted vector embedding.